Study Design and Data Analysis for Scientists

STA2007H postgraduate version

Course Information

Postgraduate students will mostly be registered for STA5014Z. This is a code that allows you to do the postgraduate version of STA2007H, without deadlines and tests.

The course is delivered in an online format, with online lectures and quizzes. Please read the description of how the course will work carefully, and post questions if you're not sure in the Q&A section on Vula.

1. Course Convenor

A/Prof Res Altwegg, Department of Statistical Sciences, Rm 6.72 PD Hahn, res.altwegg@uct.ac.za

2. Lecturers

- Dr Res Altwegg (see details above)
- Dr Birgit Erni, PD Hahn Rm 6.64, birgit.erni@uct.ac.za
- Mr Greg Distiller, PD Hahn Rm 6.67 , greg.distiller@uct.ac.za

3. Course Administrator

Contact Celene Jansen-Fielies (celene.jansen-fielies@uct.ac.za) if you have any queries about your course registration or other administrative matters. There is also a Q&A section on Vula 'General course and admin questions' which you can use to post specific admin questions during the course.

Please include your student number and the course code (STA2007H Postgraduate) in all emails.

4. Prerequisites

A pass in one of STA1000F/S, STA1007S or STA1006S, and a pass in one of MAM1004F/H, MAM1000W or MAM1005H or equivalent.

5. Course Aims:

The course aims to equip students with practical experience and skills in analyzing data, using some statistical techniques frequently used in the sciences. The skills include designing experiments, choosing appropriate statistical methods for visual display and statistical modelling of data, model checking, interpretation and reporting of statistical results, and understanding limitations of statistical methods and data. By the end of the course a student should have gained enough confidence to transfer these skills to new problems or data sets in their own profession.

6. Course Outline

Module 0: Introduction to course and statistical modelling

Module 1: Regression

- Work unit 1: Correlation and simple regression
- Work unit 2: Multilinear regression
- Work unit 3: Extensions of the linear model
- Work unit 4: Model Selection

Module 2: Design and analysis of experiments

- Work unit 1: Introduction to experimental design
- Work unit 2: Completely randomized designs
- Work unit 3: Randomized Block Designs
- Work unit 4: Factorial Experiments

Module 3: Generalized linear models

- Work unit 1: Introduction to generalized linear models
- Work unit 2: Logistic regression
- Work unit 3: Poisson regression

7. Learning Outcomes

At the end of this course the student should be able to

- identify appropriate methods to summarize and present data
- identify appropriate statistical methods and models to analyze data
- understand the issues involved in planning experiments (controls, blocks, replication, randomization)
- design a basic experiment
- implement multiple regression, analysis of variance and GLMs using statistical software
- report results and draw conclusions from statistical output
- understand the limitations of particular statistical methods and a given data set
- present statistical results in a short report
- describe in a short report statistical methods used for a project/assignment
- judge whether a particular model is adequate/appropriate for a given data set, using model checking procedures
- compare different models using model selection criteria
- understand the differences between normal, binary, proportional and count data and why different models are required for these
- criticise an experimental design
- understand the basics of statistical models
- interpret output/results from statistical software packages
- understand statistical concepts such as interactions, confounding, standard errors, confidence intervals, etc.
- use the above terminology in discussing statistical results
- use statistical software for exploratory data analysis, plots, linear regression, analysis of variance, and generalized linear models
- use statistical software for basic data manipulation, and some basic simulations
- formulate statistical hypotheses
- formulate statistical models using appropriate notation

8. Course Components:

The course has three main modules (plus a short introductory module). Modules 1 to 3 have multiple work units.

The schedule for undergraduate students is available on Vula. You can use this as a guideline for how much time you should spend on each section. We expect that the total amount of time you need to spend on the course material amounts to at least 120 hours (10 hours per week for a 12 week semester).

There will be no deadlines for quizzes, assignments and projects, except that these need to be completed to satisfaction before the exam can be written.

1. **Readings:** There is no prescribed textbook. All notes are provided on Vula in the form of pdf files. At the beginning of each section you will be required to read up on the sections in these notes. Because the videos are short, they cover only the key concepts. The notes complement the videos and go into much greater detail than the videos. Sections covered in the notes but not in the videos are therefore also important and examinable.

2. **Videos:** These complement the notes. They are short and only go into the key concepts.

   ** *Watching the videos through Vula while on campus will not impact on your internet data limit set by UCT, but watching from home will incur data downloading costs. If you are planning to watch the videos off campus, you could download them while at UCT and watch them offline so as to reduce connectivity problems.*

3. **Quizzes:** Many of the reading and video sections are followed by a quiz. All quizzes are compulsory. The quizzes mainly aim to immediately deepen your knowledge after a reading or a video. Through immediate feedback you can best see for yourselves if you have grasped the ideas just covered, and discover problem areas. You can repeatedly submit the quizzes until you get them right, until the deadline. An average of 80% for the quizzes is a DP requirement. Quizzes do not count towards the class record (but are essential for DP). All quizzes must be submitted via Vula.

4. **Assignments:** Roughly each work unit has an assignment, roughly one assignment per week. Assignment submissions are via Assignment quizzes. These are accessible either from the module pages or directly from the Tests & Quizzes Tab on Vula. Assignments will mostly require you to use R and the new statistical methods on a new data set. Assignments count towards your class record.

5. Short introductory **R workshops** will be held during the beginning of each semester. See Vula announcements for more detail.

Assessment:

Two projects and the assignments will count equally towards the class record (one third each). Class record counts 24% of final mark. One 2-hour paper written in June / November counts 76%. A subminimum of 40% is required in examinations.

DP Requirements:

At least 35% for class record and satisfactory completion of the projects (subminimum of 40%), and 80% average across all quizzes. Once the DP lists have been published you have 5 days to appeal and check year marks, after which your year mark will be made final.

Plagiarism:

Plagiarism will not be tolerated. Please familiarise yourself with what plagiarism is and include a plagiarism statement with all work submitted for assessment.
http://www.uct.ac.za/downloads/uct.ac.za/about/policies/plagiarism_students.pdf